Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation [preprint]

Chromosome-level assembly of the Atlantic silverside genome reveals extreme levels of sequence diversity and structural genetic variation

Anna Tigano[1,2], Arne Jacobs[1], Aryn P. Wilder[1,3], Ankita Nand[4], Ye Zhan[4], Job Dekker[4,5], Nina O. Therkildsen[1]

[1]Department of Natural Resources, Cornell University, Ithaca, NY, USA

[2]Department of Molecular, Cellular and Biomedical Sciences, University of New Hampshire, Durham, NH, USA

[3]Conservation Genetics, San Diego Zoo Global, Escondido, CA, USA

[4] Program in Systems Biology, University of Massachusetts Medical School, Worcester, MA 01605, USA

[5] Howard Hughes Medical Institute, Chevy Chase, MD 20815, USA

Abstract

The levels and distribution of standing genetic variation in a genome can provide a wealth of insights about the adaptive potential, demographic history, and genome structure of a population or species. As structural variants are increasingly associated with traits important for adaptation and speciation, investigating both sequence and structural variation is essential for wholly tapping this potential. Using a combination of shotgun sequencing, 10X Genomics linked reads and proximity-ligation data (Chicago and Hi-C), we produced and annotated a chromosome-level genome assembly for the Atlantic silverside (*Menidia menidia*) - an established ecological model for studying the phenotypic effects of natural and artificial selection - and examined patterns of genomic variation across two individuals sampled from different populations with divergent local adaptations. Levels of diversity varied substantially across each chromosome, consistently being highly elevated near the ends (presumably near telomeric regions) and dipping to near zero around putative centromeres. Overall, our estimate of the genome-wide average heterozygosity in the Atlantic silverside is the highest reported for a fish, or any vertebrate, to date (1.32-1.76% depending on inference method and sample). Furthermore, we also found extreme levels of structural variation, affecting ~23% of the total genome sequence, including multiple large inversions (> 1 Mb and up to 12.6 Mb) associated with previously identified haploblocks showing strong differentiation between locally adapted populations. These extreme levels of standing genetic variation are likely associated with large effective population sizes and may help explain the remarkable adaptive divergence among populations of the Atlantic silverside.

Introduction

Standing genetic variation is widely recognized as the main source of adaptation (Barrett & Schluter 2008; Tigano & Friesen 2016) and is important for natural populations to maximize their potential to adapt to changes in their environment. As genetic diversity is the result of the interplay of mutation, selection, drift and gene flow, the levels and patterns of standing genetic variation found within a species can provide important insights not only about its adaptive potential but also about its demographic and evolutionary history.

Traditionally, quantification of standing genetic variation has been based on sequence variation, often across a limited number of genetic markers, or small microsatellite repeats. As an increasing number of empirical studies shows the mosaic nature of the genome (Pääbo 2003) with different genomic regions showing vastly different levels of diversity and differentiation (e.g., Martinez Barrio et al. 2016; Campagna et al. 2017; Murray et al. 2017; Sardell et al. 2018), it is evident that small marker panels do not grant the resolution to describe variation in diversity across the genome (Dutoit et al. 2016). Furthermore, structural variation, including changes in the position, orientation, and number of copies of DNA sequence, is generally neglected as a type of standing genetic variation. Structural variation has been associated directly or indirectly with many traits involved in speciation and adaptation and is abundant in the few genomes in which they have been catalogued (Wellenreuther & Bernatchez 2018; Catanach et al. 2019; Lucek et al. 2019; Mérot et al. 2020; Tigano et al. 2020; Weissensteiner et al. 2020). Structural variants can directly affect phenotypic traits, such as the insertion of a repeated transposable element in the iconic case of industrial melanism in the peppered moth (*Biston betularia*; Van't Hof et al. 2016), or may promote the maintenance of divergent haplotypes between locally adapted populations or groups (e.g. ecotypes or morphs) within single populations via recombination suppression (e.g., Faria et al. 2019; Kess et al. 2020). Structural variation is therefore a key source of standing genetic variation, which can also play an important role in rapid evolutionary responses to environmental change (Reid et al. 2016). To better assess levels of standing variation and understand how demographic and evolutionary factors contribute to their distribution in the genome, we need to examine large proportions of the genome, preferably its entirety, and examine sequence and structural variation jointly. A high-quality reference genome for the species of interest is therefore fundamental as we need both broad coverage to accurately assess variation in levels of standing sequence variation across the genome, and high contiguity to investigate standing structural variation.

The Atlantic silverside (*Menidia menidia*), a small coastal fish distributed along the Atlantic coast of North America, shows a remarkable degree of local adaptation in a suite of traits, including growth rate, number of vertebrae, and temperature-dependent sex determination (Hice et al. 2012), that are associated with strong environmental gradients across its wide latitudinal range. This species also provided the first discovery of temperature-dependent sex determination in fishes (Conover & Kynard 1981) and was one of the first species in which countergradient phenotypic variation was documented (Conover & Present 1990). Through extensive prior work, the Atlantic silverside has, in fact, become an important ecological model to study the phenotypic effects of selection, both natural and artificial, in the wild and under controlled conditions in the lab (Conover & Munch 2002; Conover et al. 2005; Hice et al. 2012). In one iconic experiment, wild-caught Atlantic silversides were subjected to different size-selective regimes to investigate the potential of fisheries to induce evolutionary change in harvested species (Conover & Munch 2002). Seventeen years later, genomic analysis of fish from that experiment identified substantial allele frequency shifts associated with rapid phenotypic shifts in growth rates (Therkildsen et al. 2019). In the absence of a reference genome, genomic reads were mapped to the silverside reference transcriptome, so only protein-coding regions of the genome were analyzed ('in-silico' exome capture). Yet, anchoring the transcriptome contigs to the medaka (*Oryzias latipes*) chromosome-level reference genome revealed that the most conspicuous allele frequency shifts clustered into a single block on chromosome 24, where more than 9,000 SNPs in strong linkage disequilibrium (LD) increased from low (< 0.05) to high frequency (~0.6) in only five generations. Additional data from natural populations across the geographical distribution of the species showed that this same block, likely spanning several Mb of the chromosome, was fixed for opposite haplotypes among wild silverside populations that naturally differ in growth rates (Conover & Present 1990; Conover & Munch 2002; Therkildsen et al. 2019). Moreover, three additional blocks comprising hundreds of genes in high linkage disequilibrium (LD) were found to be segregating among the natural populations, with each LD block ('haploblocks' hereafter) mapping predominantly to unique medaka chromosomes (Wilder et al. 2020). Similar to the haploblock on chromosome 24, opposite haplotypes in these haploblocks were nearly fixed between natural populations that otherwise showed low genome-wide pairwise differentiation. Furthermore, strong LD between genes in these blocks suggested that local recombination suppression, possibly due to inversions, and natural selection maintained these divergent haploblocks in the face of gene flow. It thus appears that large haploblocks play an important role in maintaining local adaptations in the Atlantic silverside, although the exact extent of the genome spanned by these haploblocks and the genomic mechanism maintaining LD are unknown.

Given the wealth of ecological information available for the Atlantic silverside and its potential as an evolutionary model to study adaptation and fishery-induced evolutionary change, developing genomic resources for this species is timely and holds great potential for addressing many pressing questions in evolutionary and conservation biology. Previous population genomic analyses based on the transcriptome reference anchored to the medaka genome were limited to the coding genes and, given the unknown degree of synteny conservation between the Atlantic silverside and the medaka, how variants relevant to adaptation and fishery-induced selection clustered in the genome was uncertain. To enable analysis of both coding and non-coding regions, to accurately estimate levels and the genomic distribution of standing genetic variation, both sequence and structural, and to reconstruct the specific genomic structure of the Atlantic silverside genome, we produced a chromosome-level genome assembly for the species using a combination of genomic approaches. Because of known geographic differentiation, we estimated levels of sequence variation within genomes from both the southern and northern parts of the distribution and characterized standing structural variation between these two genomes. Finally, we tested whether the haploblocks identified on four different chromosomes between southern and northern populations were associated with large inversions as the patterns of differentiation and LD suggested (Therkildsen et al. 2019). Our work illustrates the wealth of information that can be obtained from the analysis of one or two genomes in the presence of a high quality reference sequence, and shows that, to the best of our knowledge, the Atlantic silverside has the highest nucleotide diversity reported for a vertebrate to date, and extreme levels of structural variation between two locally adapted populations. The distribution of diversity across the genome is strongly affected by structural variants and, seemingly, by genome features such as centromeres and telomeres. These results taken together highlight the importance of high-quality genomic resources as they enable the joint analysis of sequence and structural variation at the whole-genome level.

Methods

Reference genome assembly

We built a reference genome for the Atlantic silverside through three steps: First, we created a draft assembly using 10X Genomics linked-reads technology (10X Genomics, Pleasanton, CA, USA); second, we used proximity ligation data - Chicago® (Putnam et al. 2016) and Dovetail™ Hi-C (Lieberman-Aiden et al. 2009) - from Dovetail Genomics to increase contiguity, break up mis-joins, and orient and join scaffolds into chromosomes; and finally, we used short-insert reads to close gaps in the scaffolded and error-corrected assembly. The data were generated from muscle tissue dissected from two lab-reared F1 offspring of Atlantic silversides collected from the wild on Jekyll Island, Georgia, USA (N 31.02, W 81.43; the southern end of the species distribution range) in May 2017. For 10X Genomics library preparation, we extracted DNA from fresh tissue from one individual using the MagAttract HMW DNA Kit (Qiagen). Prior to library preparation, we selected fragments longer than 30 kb using a BluePippin device (Sage Science). A 10X Genomics library was prepared following standard procedure and sequenced using two lanes of paired-end 150 bp reads on a HiSeq2500 (rapid run mode) at the Biotechnology Resource Center Genomics Facility at Cornell University. To assemble the linked reads, we ran the program *Supernova* (Weisenfeld et al. 2017) from 10X Genomics with varying numbers of reads and compared assembly statistics to identify the number of reads that resulted in the most contiguous assembly. Tissue from the second individual was flash-frozen in liquid nitrogen and shipped to Dovetail Genomics, where Chicago and Hi-C libraries were prepared for further scaffolding. These long-range libraries were sequenced in one lane of Illumina HiSeq X using paired-end 150 bp reads. Two rounds of scaffolding with *HiRise*™, a software pipeline developed specifically for genome scaffolding with Chicago and Hi-C data, were run to scaffold and error-correct the best 10X Genomics draft assembly using Dovetail long-range data. Finally, the barcode-trimmed 10X Genomics reads were used to close gaps between contigs.

For each of the intermediate and the final assemblies we produced genome contiguity and other assembly statistics using the *assemblathon_stats.pl* script from the Korf Laboratory (https://github.com/KorfLab/Assemblathon/blob/master/assemblath on_stats.pl) and assessed assembly completeness with *BUSCO v3* (Simão et al. 2015) using the Actinopterygii gene set (4584 genes).

We estimated the genome size and heterozygosity (i.e. the nucleotide diversity π within a single individual) from the raw 10X Genomics data using a k-mer distribution approach. We removed barcodes with the program *longranger basic*, trimmed all reads to the same length of 128 bp (as read length is in the equation to estimate genome size) with *cutadapt* (Martin 2011), and estimated the distribution of 25-mers using *Jellyfish* (Marçais & Kingsford 2011). Finally, we analyzed the 25-mers distribution with the web application of *GenomeScope* (Vurture et al. 2017), which runs mixture models based on the binomial distributions of k-mer profiles to estimate genome size, heterozygosity and repeat content.

Synteny with medaka

The chromosome-level genome assembly of medaka (*Oryzias latipes*) was used by Therkildsen et al. (2019) to order and orient contigs of the Atlantic silverside transcriptome (Therkildsen & Baumann 2020). Although the two species carry the same number of chromosomes (Uwa & Ojima 1981; Warkentine et al. 1987) and few interchromosomal rearrangements have been observed between other species within the Atherinomorpha clade (Amores et al. 2014; Miller et al. 2019), the estimated divergence time between medaka and Atlantic silverside is 91 million years (estimate based on 15 studies, timetree.org) and the degree of syntenic conservation between the two species was unknown. We assessed synteny between the two species using the newly assembled Atlantic silverside reference genome. We aligned the silverside genome to the medaka genome (GenBank assembly accession GCA_002234675.1) with the *lastal* program in *LAST* (Kiełbasa et al. 2011; Frith & Kawaguchi 2015) using parameters optimized for distantly related species (*-m100 -E0.05*). Given the deep divergence between the two species, we kept low-confidence alignments (*last-split -m1*). We filtered alignments shorter than 500 bp and visualized syntenic relationships only for silverside scaffolds longer than 1 Mb (‘chromosome assembly’, see below) using the software CIRCA (omgenomics.com/circa).

Repeat and gene annotation

We annotated the Atlantic silverside genome using a combination of the *BRAKER2* (Hoff et al. 2019) and *MAKER* (Holt & Yandell 2011) pipelines, which combine repeat masking, *ab initio* gene predictor models and protein and transcript evidence for *de novo* identification and annotation of genes. To annotate repetitive elements, we first identified repeats *de novo* in the Atlantic silverside genome using *Repeatmodeler* (Smit & Hubley 2008) and NCBI as a search engine and combined the resulting species-specific library with a library of known repeats in teleosts (downloaded from the RepBase website (Bao et al. 2015) in July 2018). The merged libraries were then used to annotate repeats in the Atlantic silverside genome with *Repeatmasker* (Smit et al. 2015). We then filtered annotated repeats to only keep complex repeats for soft-masking. Next, we used *BRAKER2* to train *AUGUSTUS* (Stanke et al. 2006; Stanke et al. 2008; Buchfink et al. 2015) on the soft-masked genome with unpublished mRNA-seq evidence from 24 Atlantic silverside individuals from different populations and developmental stages, along with protein homology evidence from six different teleost species (medaka [*Oryzias latipes*], tilapia [*Oreochromis aureus*], platyfish [*Xiphophorus maculatus*], zebrafish [*Danio rerio*], stickleback [*Gasterosteus aculeatus*] and fugu [*Takifugu rubripes*]), which were downloaded from ensemble.org (Ensembl 98; Cunningham et al. 2019) and the UniProtKB (Swiss-Prot) protein database. Second, we ran five rounds of annotation in *MAKER* using different input datasets. The first round of *MAKER* was performed on the genome with only complex repeats masked using the non-redundant transcriptome of Atlantic silverside (Therkildsen and Palumbi 2017, Therkildsen and Baumann 2020) as mRNA-seq evidence, and the six protein sequence datasets from other species as protein homology evidence. We then trained *SNAP* (Korf 2004) on the output of the initial *MAKER* run for *ab initio* gene model prediction. We ran *MAKER* a second time adding the SNAP *ab initio* gene predictions. Using the *MAKER* output from this second round, we re-trained *SNAP* and ran *MAKER* three additional times (round 3 to 5) including the updated *SNAP* gene predictions, the *AUGUSTUS* gene predictions from *BRAKER2* and the updated *MAKER* annotation.

Lastly, we performed a functional annotation using *Blast2GO* in *Omnibox v.1.2.4* (Götz et al. 2008) utilizing the UniProtKB (Swiss-Prot) database and *InterProScan2* results. Annotated Atlantic silverside nucleotide sequences for all predicted genes were blasted against the UniProtKB database using *DIAMOND* v. 0.9.34 (Buchfink et al. 2015) with an e-value cutoff of 10^{-5}. *InterProScan2* was used to annotate proteins with *PFAM* and *Panther* annotations and identify GO terms. *Blast2GO* default mapping and annotation steps were performed using both lines of evidence to create an integrated annotation file.

Comparison of sequence and structural standing genetic variation between populations

As Atlantic silversides from Georgia show strong genomic differentiation from populations further north, primarily concentrated in large haploblocks on four chromosomes (Therkildsen et al. 2019; Wilder et al. 2020), we also sequenced the genome of a representative individual from Mumford Cove, Connecticut, USA (N 41.32°, W 72.02°) collected in June 2016 for comparison. Genomic DNA was extracted from muscle tissue using the DNeasy Blood and Tissue kit (Qiagen) and normalized to 40 ng/µl. We prepared a genomic DNA library using the TruSeq DNA PCR-free library kit (Illumina) following the manufacturer's protocol for 550 bp insert libraries. The shotgun library was sequenced using paired-end 150 bp reads on an Illumina HiSeq4000.

We estimated genome size and heterozygosity from the raw data from this shotgun library using the same k-mer approach as for the Georgia individual described above. To compare our heterozygosity estimates in Atlantic silversides from Connecticut and Georgia with other fish species, we searched the literature for heterozygosity estimates from Genomescope with the keywords "Genomescope heterozygosity fish", or from variant calling methods in other fish genomes, using Google Scholar. We also estimated heterozygosity directly by calculating the proportion of heterozygous sites in each genome. For the Georgia individual we used the processed 10X data as above. For the Connecticut individual we trimmed adapters and low-quality data from the raw shotgun data in *Trimmomatic* (Bolger et al. 2014). We mapped data from the two libraries to the chromosome assembly (only the largest 27 scaffolds - see Results) with *bwa mem* (Li & Durbin 2009) and removed duplicates with *samblaster* (Faust & Hall 2014). We called variants with *bcftools mpileup* and *bcftools call* (Danecek et al. 2014). As areas of the genome covered by more than twice the mean sequencing depth could represent repetitive areas or assembly artefact, we calculated genome coverage for each of the two libraries with *genomeCoverageBed* from *BEDtools* (Quinlan & Hall 2010) and identified the depth mode from the calculated distribution (95x for the southern genome and 74x for the northern genome). We then filtered variants that were flagged as low-quality, that had mapping quality below 20, sequencing depth below 20, and more than twice the mode sequencing depth for each of the two libraries using *bcftools filter* (Li et al. 2009). To accurately estimate the proportion of heterozygous sites in the genome, we subtracted the number of sites that had sequencing depth below 20 and above twice the mode sequencing depth from the total genome size (to get the sum of sites that could be identified as either homozygous or heterozygous based on our criteria). To visualize variation along the genome, we plotted estimates of heterozygosity in 50-kb sliding windows along the genome for each of the two individuals using the *qqman* package (Turner 2014) in R (R Core Team 2019). To assess the reduction in diversity in protein-coding regions due to positive and purifying selection, we calculated heterozygosity in the regions annotated as coding sequences only and compared this to the genome-wide estimate.

Finally, we identified structural variants (SVs) segregating between the Connecticut and Georgia genomes using *Delly2 v.0.8.1* (Rausch et al. 2012). For this analysis we used the shotgun library data (74x coverage) from Connecticut mapped to the Georgia reference genome as described above. We called SVs using the command *delly call* and default settings. As genotyping a single individual in *Delly* is prone to false positives we applied the following stringent filters: We retained only homozygous SVs (*vac=2*) that passed quality filters (*PASS*) and that had at least 20 reads supporting the variant calls, whether they came from paired-end clustering or split-read analysis or a combination of the two, but not more than 100 reads since these could be due to repetitive elements in the genome. As *Delly2* outputted redundant genotypes, e.g. inversions that had slightly different breakpoints were reported as independent variants, we used *bedtools merge* to merge these overlapping features. To validate duplication calls we also calculated coverage for each of these variants and retained only those putative duplications that had coverage more than 1.8-fold the whole genome sequencing depth (74x).

To confirm the large SVs observed between the two genomes examined, we generated a second Hi-C library from an Atlantic silverside individual caught in Mumford Cove, Connecticut in June 2016 (different from the sample used for the shotgun assembly). Liver tissue was excised and digested for 2 hours in collagenase digestion buffer (perfusion buffer plus 12.5 µM CaCl2 plus collagenases II and IV (5 mg/ml each)). The cell suspension was then strained through a 100 µm cell strainer, washed with 1 ml cold PBS three times, resuspended in 45 ml PBS, and quantified in a hemocytometer. The cross-linking protocol was modified from Belton et al. (2012) as follows. 1.25 ml of 37% formaldehyde was added twice to the cell preparation, then incubated at room temperature for 10 minutes, inverting every 1-2 minutes. To quench the formaldehyde in the reaction, 2.5 ml of 2.5 M glycine was added three times. The sample was incubated at room temperature for 5 minutes, then on ice for 15 minutes to stop the cross-linking. The cells were pelleted by centrifugation (800g for 10 min), and the supernatant was removed. The sample thus obtained was flash frozen in liquid nitrogen and stored at -80°C. Hi-C library preparation was performed as described previously (Belaghzal et al. 2017), except that ligated DNA size selection was omitted. 50 million fish liver cells were digested with *DpnII* at 37°C overnight. DNA ends were filled with biotin-14-dATP at 23°C for 4 hours. DNA was then ligated with T4 DNA ligase at 16°C overnight. Proteins were removed by treating ligated DNA with proteinase-K at 65°C overnight. Purified, proximally ligated molecules were sonicated to obtain an average fragment size of 200 bp. After DNA end repair, dA-tailing and biotin pull down, DNA molecules were ligated to Illumina TruSeq sequencing adapters at room temperature for 2 hours. Finally, the library was PCR-amplified and finalized following the Illumina TruSeq Nano DNA Sample Prep kit manual. Paired-end 50 bp sequencing was performed on a HiSeq4000.

The two Hi-C libraries from Connecticut and Georgia (the latter prepared by Dovetail) were mapped to the Atlantic silverside chromosome assembly using the *Distiller* pipeline (github.com/mirnylab/distiller-nf). Interaction matrices were binned at 50 and 100 kb resolution and intrinsic biases were removed using the Iterative Correction and Eigenvector decomposition (ICE) method (Imakaev et al. 2012). Large inversions (> 1 Mb) were identified by visual inspection of Hi-C maps as discontinuities that would be resolved when the corresponding section of the chromosomes were to be inverted (Dixon et al. 2018; Corbett-Detig et al. 2019)**.** These discontinuities generate a distinct "butterfly pattern" with signals of more frequent Hi-C interactions where the projected coordinates of the breakpoints meet.

Results

Genome assembly and assessment of completeness

We obtained the best draft assembly (with the highest contiguity; N50 = 1.3 Mb) from the 10X data when we used 270 million reads as input to *Supernova*. Contiguity increased more than 2-fold with Dovetail Chicago data (scaffold N50 = 2.9 Mb) and more than 10-fold with Dovetail Hi-C data (scaffold N50 = 18.2 Mb). Summary statistics for each of the intermediate genome assemblies (10X, Dovetail Chicago, and Dovetail Hi-C) are presented in Table 1. The final assembly – including scaffolds longer than 1 kb only – was 620 Mb in total length. Overall, this assembly showed high contiguity, high completeness and a low proportion of gaps (Table 1).

Analysis of the presence of BUSCO genes showed that only 5.9% of the Actinopterygii gene set were missing from the assembly. Although the number of missing genes did not decrease dramatically from the 10X assembly to the final assembly (from 6.6 to 5.9%), the addition of proximity ligation data (Chicago and Hi-C) increased the number of complete genes (from 88.1 to 89.6%) and decreased the number of duplicated (from 4.1 to 2.9%) and fragmented genes (from 5.3 to 4.5%). Contiguity did not come at the cost of increased gappiness, as stretches of N's made up only 3% of the final assembly. The reduction of the assembly to its longest 27 scaffolds ('chromosome assembly'- a 25% reduction in sequence) increased missing genes by only 3.1% and reduced duplicated genes to 1.9%. K-mer analyses based on raw data from the reference genome estimated a genome size of 554 Mb, 76 Mb shorter than the final assembly and 88 Mb longer than the chromosome assembly.

Synteny with Medaka

The alignment of the 27 largest Atlantic silverside scaffolds to the medaka genome revealed a high degree of synteny conservation, especially considering the evolutionary distance between the two species. Each Atlantic silverside scaffold mapped mostly to only one medaka chromosome, and 22 of the 24 medaka chromosomes had matches with only one Atlantic silverside scaffold each (Fig. 1). Two medaka chromosomes, 1 and 24, had matches with three and two silverside scaffolds, respectively (Fig. 1). Based on these results, karyotype data confirming that the medaka and silverside have the same number of chromosomes (Uwa & Ojima 1981; Warkentine et al. 1987), and additional support from the Hi-C data from the Connecticut individual, we ordered and renamed the Atlantic silverside scaffolds according to the orthologous medaka chromosomes. We grouped the three and two scaffolds that mapped to medaka chromosomes 1 and 24, respectively, into one pseudo-chromosome each and renamed them accordingly. Although we did not observe large interchromosomal rearrangements in the alignment of the silverside and medaka genomes (Fig. 1), intrachromosomal rearrangements were common (Fig. 1; Fig. S1). The most conspicuous chromosomal rearrangements were large inversions, intrachromosomal translocations and duplications (Fig. 1; Fig. S1). On chromosomes 8, 11, 18 and 24, where large geographically differentiated haploblocks were identified among natural silverside populations, several translocations and inversions were evident, indicating poor intrachromosomal synteny (Fig. 1). This was also the case for most of the other chromosomes (Fig. S1).

Repeat and gene annotation

The identified repetitive elements made up 17.73% of the Atlantic silverside genome, in line with expectations based on fish species with similar genome sizes (Yuan et al. 2018). The biggest proportion of these repeats was made up of interspersed repeats (15.34% of the genome), while transposable elements constituted 8.83% of the genome overall (0.90% of SINEs, 2.79% of LINEs, 1.54% of LTR elements, and 3.60% of DNA elements). Our gene prediction pipeline identified a total of 21,644 protein coding genes, a number consistent with annotated gene counts in other fish species (Lehmann et al. 2019; Ozerov et al. 2018). Analysis in *Blast2GO* based on homology and *InterProScan2* resulted in functional annotation of 17,602 out of the 21,644 protein coding genes (81.3%; https://github.com/atigano/Menidia_menidia_genome/annotation/). Further, *InterProScan2* detected annotations (*Panther* or *PFAM*) for an additional 1,511 genes, for which no BLAST results were obtained.

Sequence and structural standing variation

K-mer analyses based on raw data resulted in similar estimates of genome sizes and levels of heterozygosity in the two samples from Georgia and Connecticut: genome size estimates differed by 20 Mb (554 Mb and 535 Mb in the Georgia and Connecticut individual, respectively) and heterozygosity estimates differed by 0.09% (1.76% and 1.67% in Georgia and Connecticut, respectively). Direct estimates of heterozygosity, i.e. based on the number of called heterozygous sites in the genome, were slightly lower and differed by 0.14% between individuals (1.32% and 1.46% in Georgia and Connecticut, respectively). Together, these estimates concordantly indicate that standing sequence variation in this species is very high (Kajitani et al. 2014), with 1 in every ~66 bp being heterozygous within each individual. These heterozygosity estimates are higher than all comparable estimates reported for other fish species, though of similar magnitude to the European sardine and two eel species (Table 2). Heterozygosity varied substantially across the genome. Within each chromosome, heterozygosity was highest toward the edges of each chromosome, presumably in areas corresponding to telomeres, decreased towards the center in a U-shape fashion, and showed a deep dip in which the number of heterozygous sites approached zero, consistent with the location of putative centromeres (Fig. 2b). Additionally, the proportions of variable sites in coding regions was ~50% of whole genome level estimates (0.68% and 0.70% in Georgia and Connecticut, respectively). Swaths of low heterozygosity were particularly evident on chromosomes 18 and 24, two of four chromosomes with highly differentiated haploblocks (Fig. 2a,b).

We identified a total of 4,900 SVs - including insertions, deletions, duplications and inversions (Supplementary File) - between the reference genome generated from Georgia samples and the re-sequenced individual from Connecticut. *Delly2* indicated that insertions were small (42-83 bp) and affected a negligible proportion of the genome, while deletions were larger and more abundant, covering 15% of the genome sequence. As an insertion in one genome corresponds to a deletion in the other genome depending on which individual is used as reference, the discrepancy between insertions and deletions is an artefact of mapping short-read sequences to a single reference, i.e. inserted sequences found only in Connecticut are not present in the reference and thus are not mapped. These results highlight the difficulties in identifying insertions and estimating their sizes from short reads. Our analysis detected a small number of duplications, covering only 0.1% of the genome. In contrast, we identified 662 inversions ranging from 203 bp to 12.6 Mb in size. In total, inversions affected 109 Mb, or 23%, of the reference genome sequence. Twenty-nine inversions were larger than 1 Mb, and five larger than 5 Mb (genomic locations in Fig. 2a and in Supplementary File). *Delly2* identified large inversions (> 1 Mb) on all four chromosomes with previously identified haploblocks . The largest inversion (~12 Mb) was identified on chromosome 8; chromosome 11 had two 1.2-Mb inversions that were 7 Mb apart; chromosome 18 had a 7.4 Mb inversion and chromosome 24 had two inversions, the first one spanning 9.4 Mb and followed by another one at a distance of 76 kb, spanning 2.3 Mb (Fig. 2a).

The independent Hi-C data from Connecticut (which was not used for genome scaffolding) supported a high degree of accuracy in the overall assembly into chromosomes, as indicated by the strong concentration of data points along the diagonal rather than elsewhere in the contact maps (Fig. 3). The contact maps also readily detected large-scale inversions (> 1 Mb) between the individual from Connecticut and the reference assembly from Georgia in three of the four chromosomes with haploblocks, i.e. 8, 18, and 24 (Fig. 3, Supplementary File). The missed detection of the inversions on chromosome 11 could either be due to their relatively smaller sizes, barely exceeding the detection threshold from Hi-C data, or because both inversion orientations segregate where the Connecticut individual used for Hi-C was sampled (Wilder et al. 2020). The breakpoints of the 12.6 and 9.4 Mb inversions on chromosomes 8 and 24, respectively, matched very closely those identified by *Delly2*, although the second 2.3 Mb inversion on chromosome 24 was not supported by Hi-C data (Figs. 2a, 3, Supplementary File). On chromosome 18, Hi-C data showed a complex series of nested and/or adjacent inversions spanning ~8.8 Mb in total, in contrast with the single inversion, and ~1.3 Mb shorter, identified by *Delly2* (Figs. 2a, 3, Supplementary File). Additional large inversions were detected from the Hi-C data on chromosomes 4, 7 and 19. Of these, the inversion on chromosome 19 was not identified from the analysis of shotgun data with *Delly2*, while those on chromosome 4 and 7 were, although with only one matching breakpoint for the inversion on chromosome 4 (Figs. 2a, 3, Supplementary File). Note that the identification of SVs from shotgun and Hi-C data were carried out by two different authors, and blindly from each other.

Discussion

We generated a chromosome-level assembly of the Atlantic silverside genome by integrating long-range information from synthetic long reads from 10X Genomics, *in vitro* proximity ligation data from Chicago libraries, and Hi-C proximity ligation data from whole cells. The resulting assembly had high contiguity and completeness. Based on karyotype information (Uwa & Ojima 1981; Warkentine et al. 1987), chromosome-level synteny with medaka, and Hi-C maps we reduced the 27 largest scaffolds to 24 putative chromosomes. This chromosome assembly is 88 Mb shorter than the genome size estimated through k-mer analysis, but has a lower number of duplicated genes, and only slightly fewer missing genes than the full assembly despite a substantial reduction in total sequence. If the proportion of complete genes in the chromosome assembly is, in fact, a good proxy for genome completeness, then the scaffolds that are not placed in chromosomes are mostly sequences that are repetitive, redundant, or that should fill gaps in the assembled chromosomes.

Heterozygosity within a sequenced individual can result in alternative alleles getting assembled into distinct scaffolds, even in genomes much less heterozygous than the Atlantic silverside (Kajitani et al. 2014; Tigano et al. 2018), so we expect some redundancy in our assembly. Considering the abundance of SVs between the two sequenced individuals, structural variation also may have contributed to the high number of smaller scaffolds not included in the chromosome assembly, as heterozygous SVs are notoriously hard to assemble (Huddleston et al. 2017). Nonetheless, the Atlantic silverside genome adds to the increasing number of high-quality fish reference genome assemblies, with the sixth highest contig N50 (202.88 kb) and the sixth highest proportion of the genome contained in chromosomes (84%, based on the genome size estimate from the k-mer analysis) compared to 27 other chromosome-level fish genome assemblies (Lehmann et al. 2019).

Patterns of synteny between the Atlantic silverside and the relatively distantly related medaka are consistent with comparisons among other teleost genomes up to hundreds of millions of years diverged: rearrangements are rare among chromosomes but common within (Amores et al. 2014; Rondeau et al. 2014; Miller et al. 2019; Pettersson et al. 2019). Consistent with this, anchoring Atlantic silverside transcriptome contigs on to medaka genome enabled the identification of four large haploblocks associated with fishery-induced selection in the lab and/or putative adaptive differences in the wild (Therkildsen et al. 2019; Wilder et al. 2020). However, the high degree of intrachromosomal rearrangements between the two species, and generally among teleosts, prevented an accurate characterization of the extent of these haploblocks and the analysis of structural variation. Differentiation between the northern and southern haplotypes seemed to extend across almost the entire length of three of the four chromosomes with haploblocks when data were oriented to medaka (Therkildsen et al. 2019; Wilder et al. 2020). However, the abundant intrachromosomal rearrangements between medaka and Atlantic silverside chromosomes (Fig. 1; Fig. S1), and the detection of large inversions in each of these four chromosomes (Figs. 2a,3) suggest that differentiation is concentrated in, and possibly maintained by, these inversions, which, albeit large, do not span whole chromosomes.

Our analysis of two genomes sequenced at high coverage suggested that levels of standing genetic variation, both sequence and structural, are extremely high in the Atlantic silverside. To our knowledge, our estimates of heterozygosity in a single individual are the highest reported for any fish species to date, including those with large census population sizes (Table 2). For example, heterozygosity, which is equivalent to nucleotide diversity (π) in one individual, in one single Atlantic silverside genome was higher than, or on par with, π estimates based on 43-50 individuals of Atlantic killifish, a species considered to have ‘extreme’ levels of genomic variation with π ranging from 0.011 to 0.016 (Reid et al. 2017, 2016). Compared to other vertebrates, genome heterozygosity in the Atlantic silverside was more than double the highest estimate reported for birds (0.7% in the thick-billed murre *Uria lomvia*; Tigano et al. 2018) and higher than the population-based 0.6-0.9% estimates in the rabbit (*Oryctolagus cuniculus*), one of the mammals with the highest genetic diversity (Carneiro et al. 2014). Among a collection of genome-wide π estimates - mostly population-based - across 103 animal, plant and fungal populations or species, only three insects and one sponge had π estimates higher than the Atlantic silverside (Robinson et al. 2016 and references therein). This unusually high level of standing sequence diversity is likely due to huge population sizes with estimated N_e exceeding 100 million individuals (Lou et al. 2018), and may underpin the remarkable degree of adaptive divergence and rapid responses to selection documented for the species.

Variation in π across the genome has been associated with variation in recombination rates, with higher diversity and recombination rates in smaller chromosomes and in proximity of telomeres in fish, mammals and birds (Ellegren 2010; Murray et al. 2017; Sardell et al. 2018; Tigano et al. 2020). In the Atlantic silverside, the decrease of heterozygosity from the ends towards the center of each chromosome is consistent with decreasing recombination rates as distance from the telomeres increases (Haenel et al. 2018; Sardell et al. 2018). However, in addition to this U-shape pattern, heterozygosity shows a dramatic, narrow dip in each chromosome far from the center of chromosomes, suggesting a strong centromere effect. Although striking differences exist between sexes and across taxa, recombination is generally reduced or suppressed around centromeres (Sardell & Kirkpatrick 2020). The Atlantic silverside karyotype, with only four metacentric and 20 non-metacentric chromosomes (i.e. submetacentric, subacrocentric, and acrocentric; Warkentine et al. 1987), further supports that these dips in heterozygosity are associated with centromeres, as the non-metacentric chromosomes enable the distinction between the effect of centromeres from the effect of distance from telomeres. In forthcoming work, linkage mapping will allow us to quantify the relative effects of centromeres and telomeres on local recombination rates and ascertain whether the recombination landscape is different between sexes.

We report a 50% reduction in heterozygosity in coding sequences compared to whole genome estimates, confirming the expectation that estimates based on exome data are not representative of whole-genome levels of standing variation. Even though the magnitude of the reduction in π within coding regions is similar to levels reported in the Atlantic killifish (Reid et al. 2017) and in the butterfly *Heliconius melpomene* (Martin et al. 2016), a substantially greater reduction is seen in the collared flycatcher (86%; Dutoit et al. 2017), suggesting that the distribution of diversity in a genome, including the difference between coding and non-coding sequence, is likely idiosyncratic to the population or species examined. Once again, a paucity of data from other species prevents us from making generalizations or identifying differences on the expected reduction in diversity in coding compared to non-coding regions across taxa, while at the same time it highlights the importance of estimating and reporting basic diversity statistics for whole genome assemblies.

We identified 4,900 structural variants that survived the stringent filters applied to maximize confidence in the identified SVs and to minimize the number of false positives due to genotyping one individual only. Our estimates are likely conservative when we consider that we filtered out all heterozygous SVs, that many SVs, particularly complex ones, are hard to identify or characterize (Chaisson et al. 2019), and that we analyzed only two genomes. Nonetheless, our analyses based on shotgun data show that SVs are abundant, affect a large proportion of the genome, with inversions covering up to 23% of the genome sequence, and range in size from small (< 50 bp) to longer than 10 Mb, with many of the largest inversions further supported by independent Hi-C data. Sunflower species of the genus *Helianthus* show a similar proportion of sequence covered by inversions (22%; Barb et al. 2014), although these were detected in comparisons between species (1.5 million years diverged) rather than within species. The few studies available on other species show that structural variation tends to affect a larger portion of the genome than single nucleotide polymorphisms (SNPs), but in proportions far lower than what we report here for the Atlantic silverside. For example, structural variation, including indels, duplication and inversions, covered three times more bases than SNPs did across six individuals of Australian snapper (*Chrysophrys auratus*; Catanach et al. 2019); short indels alone affected 4% of the genome of two individuals from the same population in the cactus mouse (*Peromyscus eremicus*; Tigano et al. 2020); inversions, duplications and deletions combined affected 3.6% of the genome across 20 individuals of *Tinema* stick insects (Lucek et al. 2019); and in cod (*Gadus morhua*) inversions covered ~7.7% of the genome (Wellenreuther & Bernatchez 2018 and references therein). Although levels of structural variation in the Atlantic silverside are extreme in comparison to these studies, a direct comparison with these and other species is hampered by a paucity of data and lack of common best practices for SVs genotyping (Mérot et al. 2020): differences in sampling, approaches, data types and filtering prevent comparisons similar to those made for standing sequence variation here and in other studies (Corbett-Detig et al. 2015; Robinson et al. 2016). Given the fast rate at which high-quality reference genomes are now generated, this will hopefully start to change.

The simple and affordable strategy we adopted only requires sequencing of a single additional shotgun library prepared from a second individual - possibly from a differentiated population to capture a broader representation of intraspecific variation - and could be easily applied in other studies to start describing variation in the prevalence and genome coverage of SVs across taxa. Here, an additional Hi-C library then allowed us to discover that the putative inversion on chromosome 18 was larger than indicated by the analysis of shotgun data and was actually constituted by a combination of two or more nested inversions. The apparent discrepancy between the breakpoints of the largest inversions identified using the two data types could reflect biological variation between the individuals analyzed. Alternatively, they may be caused by the different strengths and limitations of the underlying analytical approaches, including the fact that the identification of SVs was computational from shotgun data, while it was manually curated from Hi-C data. Although the analysis of only two individuals does not capture the full spectrum of intra- and inter-population variation, integrating different approaches has allowed us to identify a set of high-confidence SVs to be validated and genotyped in a larger number of individuals with lower coverage data (Mérot et al. 2020).

The joint analysis of sequence and structural variation reveals interesting features of the previously identified haploblocks. The chromosome-level assembly of the Atlantic silverside genome a) confirms that previously identified large haploblocks (Wilder et al. 2020) are associated with inversions and allows to measure their real extent ; and b) highlights how genomic heterogeneity is multidimensional by revealing that even haploblocks showing similar patterns of differentiation can show vastly different patterns of genetic diversity. On chromosomes 18 and 24, large swaths of reduced heterozygosity (Fig. 2b) are associated with an inversion affecting the same area, which strongly indicates that the inversion promotes differentiation between genomes from Connecticut and Georgia in this region, likely through suppressed recombination. Of note, however, the segment of chromosome 24 preceding the inversion (0-722 kb) shows an even stronger reduction in heterozygosity than the adjacent inversion. While this additional reduction may be due to stronger recombination suppression in this area, the mechanism explaining this pattern remains to be investigated. In contrast, no reduction in diversity is associated with the inversion on chromosome 8 - the largest of them all (12.6 Mb) - or with the smaller inversions on chromosome 11. Such differences among haploblocks likely reflect idiosyncratic evolutionary histories and adaptive significance of the underlying inversions, whose investigation is now enabled by the chromosome-level genome assembly that we presented here. Hence, our analyses provide an empirical example of the importance of analyzing both sequence and structural variation to understand the mechanism underpinning the heterogeneous landscape of genomic diversity and differentiation.

Building on prior analysis based on in silico exome capture (Therkildsen & Palumbi 2017; Therkildsen et al. 2019; Therkildsen & Baumann 2020), this newly assembled reference genome provides an important resource for using the Atlantic silverside as a powerful model for investigating many outstanding questions in adaptation genomics, for example related to the abundance, distribution and adaptive value of structural variants; the relative role of coding and non-coding regions; the importance of sequence variation vs. structural variation in both human-induced evolution and local adaptation; and the demographic and evolutionary factors generating the genomic landscape of diversity and differentiation in this and other species.

Acknowledgements

We would like to thank Hannes Baumann for rearing and collecting fish, Nicolas Lou for assistance with sampling, Harmony Borchardt-Wier for help in the lab, Leif Andersson for discussions on levels of heterozygosity in fish and other vertebrates, Peter Schweitzer at the Cornell Biotechnology Resource Center for advice on the 10X Genomics sample preparation and analysis, and Mark Daly at Dovetail Genomics for help with generating and interpreting the proximity ligation data. This work was funded by a National Science Foundation grant to N.O.T. (OCE-1756316) and the National Human Genome Research Institute (R01 HG003143 to J.D). J.D. is an investigator of the Howard Hughes Medical Institute.

Author contributions

AT and NOT designed the study with input from JD; AJ performed the gene annotation; AW collected samples and performed lab work; AT, NOT, AW, YZ, AN and JD generated and analyzed the data; NOT and JD funded the project. AT wrote the paper with critical input from all authors.

Data accessibility

The genome assembly and associated sequence data from Georgia and the raw data from the shotgun library from Connecticut will be available under ENA accession number #######. Scripts for the genome assembly and all other analyses can be found at https://github.com/atigano/Menidia_menidia_genome/.

Tables and Figures

Table 1. Summary statistics for each of the intermediate and final assemblies produced.

	10X	Dovetail Chicago	Dovetail Hi-C	Final assembly	Chromosome assembly*
Total length	645.45 Mb	647.32 Mb	647.39 Mb	620.04 Mb	465.69 Mb
Longest Scaffold	12,248,921 bp	12,871,938 bp	26,678,928 bp	26,678,928 bp	26,678,928 bp
Number of scaffolds	99,541	80,990	80,312	42,220	27
Number of scaffolds > 1kb	61,451	42,898	42,220	42,220	27
Contig N50	39.55 kb	39.51 kb	39.51 kb	105.76 kb	202.88 kb
Scaffold L50/N50	83/1.328 Mb	42/2.936 Mb	16/18.159 Mb	15/18.199 Mb	11/19.68 Mb
% gaps	2.69%	2.97%	2.98%	3.08%	3.00%
BUSCOs** (n=4584)	C:88.1%, F:5.3%, M:6.6%	C:89.5%, F:4.6%, M:5.9%	C:89.6%, F:4.8%, M:5.6%	C:89.6%, F:4.5%, M:5.9%	C:88.3%, F:2.7%, M:9.0%

* The ‘chromosome assembly’ is the subset of scaffolds > 1 Mb from the ‘Final assembly’

** [C=complete, F=fragmented, M=missing]

Table 2. Examples of heterozygosity levels in single fish genomes, estimated either with GenomeScope from raw sequencing data or through direct calling of heterozygous sites.

Common name	**Scientific name**	**Heterozygosity**	**Method**	**Reference**
Atlantic silverside	***Menidia menidia***	**1.67-1.76%**	GenomeScope	**This study**
European sardine	*Sardina pilchardus*	1.60–1.75%	GenomeScope	Machado et al. 2018
American eel	*Anguilla rostrata*	1.5-1.6%	GenomeScope	Jansen et al. 2017
European eel	*Anguilla anguilla*	1.48-1.59%	GenomeScope	Jansen et al. 2017
Pearlscale pygmy angelfish	*Centropyge vrolikii*	1.36%	GenomeScope	Fernandez-Silva et al. 2018
Marine medaka	*Oryzias melastigma*	1.19%	GenomeScope	Kim et al. 2018
Large yellow croaker	*Larimichthys crocea*	1.06%	GenomeScope	Mu et al. 2018
Javafish medaka	*Oryzias javanicus*	0.96%	GenomeScope	Takehana et al. 2020
Greater amberjack	*Seriola dumerili*	0.65%	GenomeScope	Sarropoulou et al. 2017
Clownfish	*Amphiprion ocellaris*	0.60%	GenomeScope	Tan et al. 2018
Hilsa shad	*Tenualosa ilisha*	0.58-0.66%	GenomeScope	Mollah et al. 2019
Whitefish	*Coregonus sp. “Balchen”*	0.44%	GenomeScope	De-Kayne et al. 2020
Corkwing wrasse	*Symphodus melops*	0.40%	GenomeScope	Mattingsdal et al. 2018
Herring	*Clupea harengus*	0.32%	Variant calling	Martinez Barrio et al. 2016
Golden pompano	*Trachinotus ovatus*	0.31%	GenomeScope	Zhang et al. 2019
Coelacanth	*Latimeria chalumnae*	0.28%	Variant calling	Amemiya et al. 2013
NA	*Lucifuga gibarensis*	0.26%	GenomeScope	Policarpo et al. 2020
Eurasian perch	*Perca fluviatilis*	0.24–0.28%	GenomeScope	Ozerov et al. 2018
Atlantic cod	*Gadus morhua*	0.20%	Variant calling	Star et al. 2011
Big-eye mandarin Fish	*Siniperca knerii*	0.16%	GenomeScope	Lu et al. 2020
Threespine stickleback	*Gasteosteus aculeatus*	0.14%	Variant calling	Jones et al. 2012
Pikeperch	*Sander lucioperca*	0.14%	GenomeScope	Nguinkal et al. 2019
African arowana	*Heterotis niloticus*	0.13%	GenomeScope	Hao et al. 2020
Orange clownfish	*Amphiprion percula*	0.12%	GenomeScope	Lehmann et al. 2019
Murray cod	*Maccullochella peelii*	0.10%	GenomeScope	Austin et al. 2017
Toothed Cuban cusk-eel	*Lucifuga dentata*	0.10%	GenomeScope	Policarpo et al. 2020

Table 3. Summary of intraspecific structural variants identified in the Atlantic silverside, and their features.

SV type	Number of variants	Size range (bp)	Sequence affected (kb)	% genome affected
Insertions	299	42-83	18	<0.01%
Deletions	3905	38-9,740,501	71,754	15%
Duplications	34	110-150,263	479	0.1%
Inversions	662	203-12,585,625	109,201	23%

Figure 1. Circos plots showing synteny between the Atlantic silverside and medaka across all chromosomes in the middle and in the four chromosomes with large haploblocks on the sides. Chromosomes are color-coded consistently among plots and the colored portion of the smaller plots refer to the medaka sequences, while the grey portion to the Atlantic silverside sequences. Alignments shorter than 500 bp were excluded. Fig. S1 shows plots for the remaining chromosomes. Note that the consistently shorter length of the Atlantic silverside genome is consistent with a lower overall estimate of genome size (554 Mb based on k-mer analysis compared to the 700 Mb of the assembled medaka genome). The three and two scaffolds making up chromosomes 1 and 24, respectively, are represented separately here and denoted by small letters.

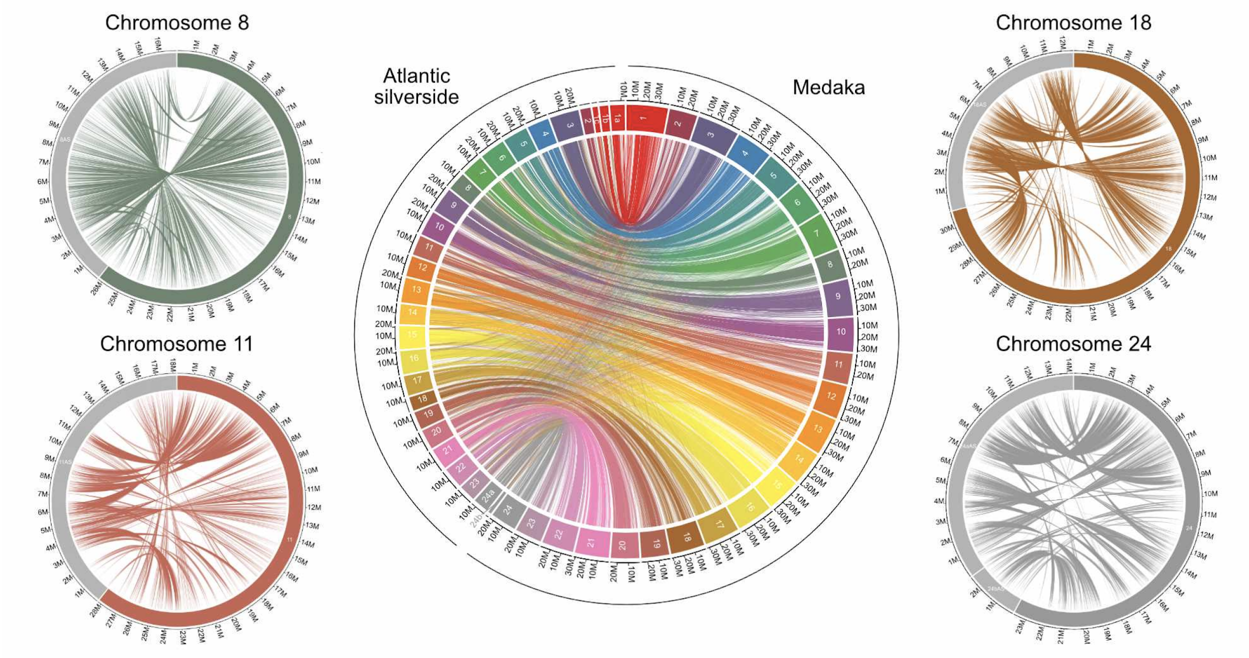


Figure 2. The genomic landscape of structural and sequence variation in Connecticut and Georgia. a) Panel showing large inversions (> 1 Mb) as identified from shotgun and Hi-C data from an individual from Connecticut mapped to the reference genome from Georgia. b) Manhattan plots showing the genomic landscape of variation in heterozygosity in 50 kb moving windows across single genomes from Connecticut and Georgia. The three and two scaffolds making up chromosomes 1 and 24, respectively, are represented separately here and denoted by small letters (e.g., 1a and 24a).

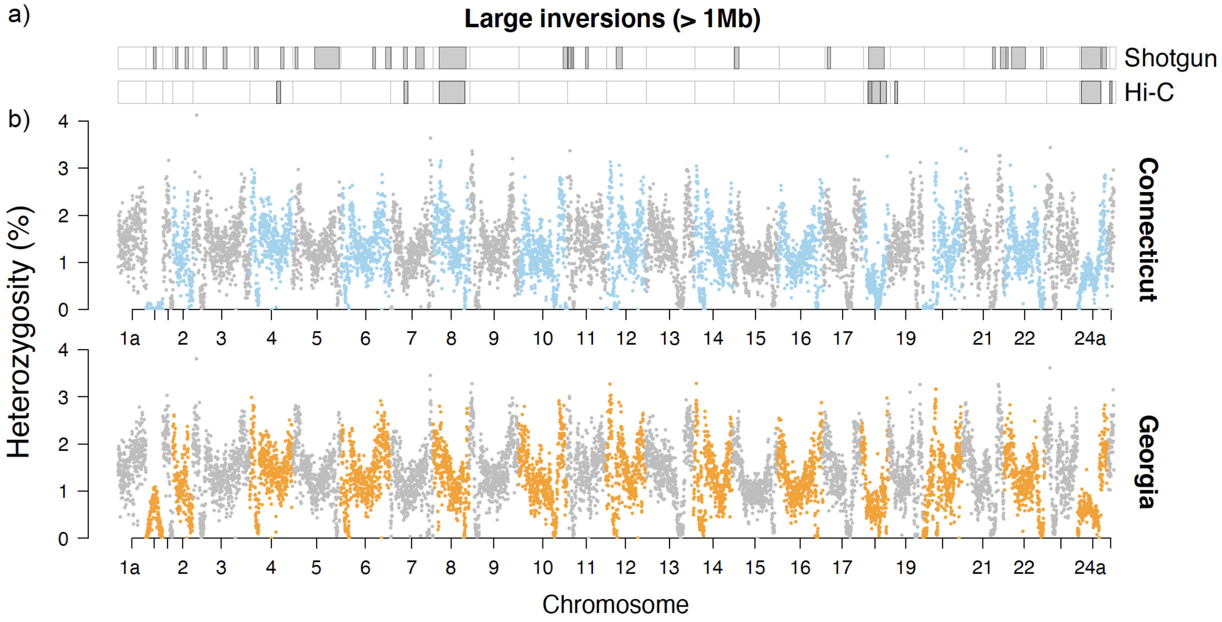
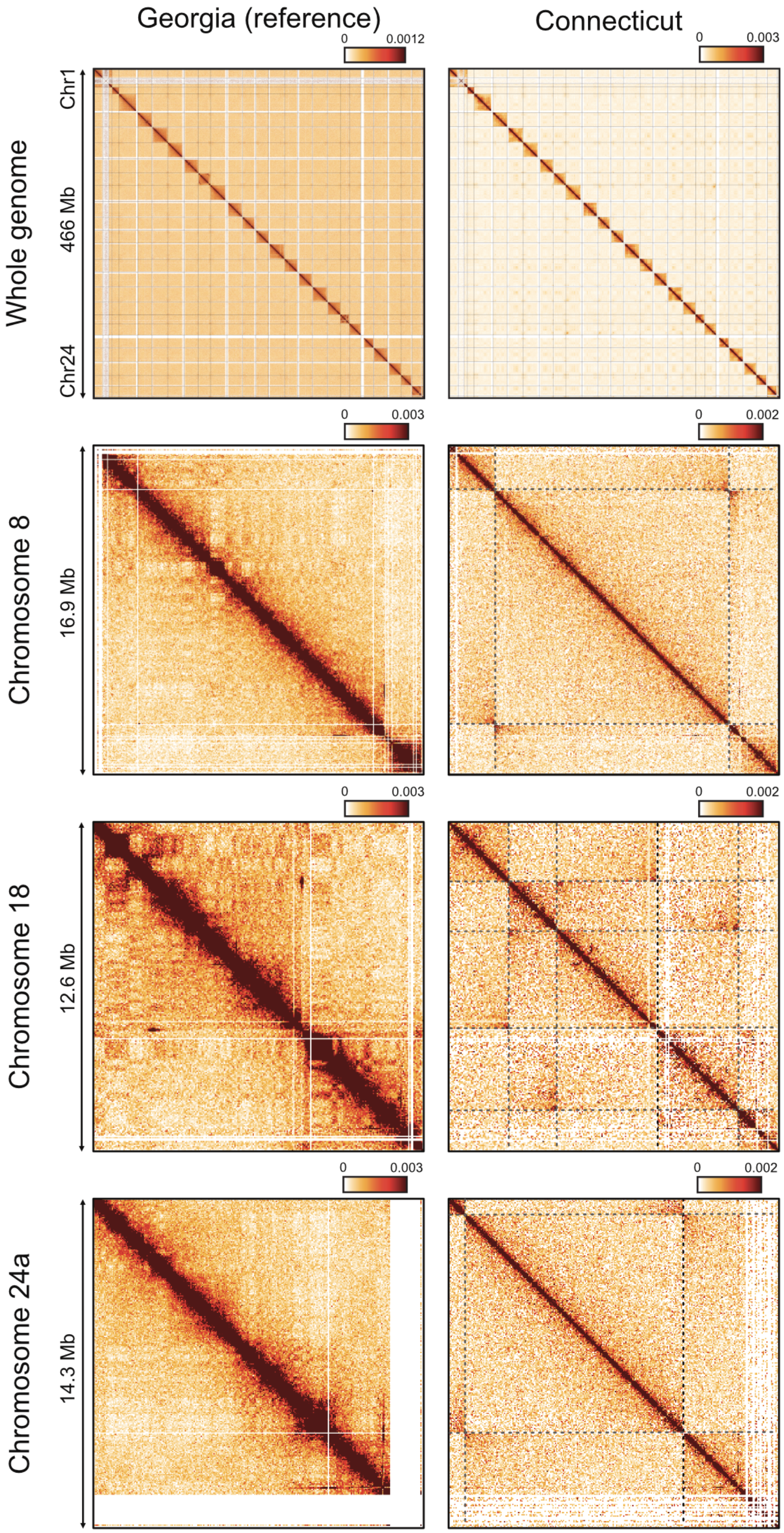


Figure 3. Hi-C contact maps of data mapped to the chromosome assembly from Georgia. Maps on the left show Hi-C data obtained from the same Georgia individual used to generate the reference assembly (mapped to self), maps on the right show data obtained from a Connecticut individual. Maps in the top panel show data for all the chromosomes binned in 100 kb sections. The three lower panels show data binned in 50 kb sections from each of the three chromosomes showing both large haploblocks in Wilder et al. (2020) and evidence for the presence of inversions from Hi-C data. Dark shades on the diagonal are indicative of high structural similarity between the reference and the Hi-C library analyzed. Dashed lines represent putative inversion breakpoints. The "butterfly pattern" of contacts observed at the point when the dashed lines meet is diagnostic of inversions.


Georgia (reference)
Connecticut
0
0.0012
0
0.003
Whole genome
Chr1
466 Mb
Chr24
0
0.003
0
0.002
Chromosome 8
16.9 Mb
0
0.003
0
0.002
Chromosome 18
12.6 Mb
0
0.003
0
0.002
Chromosome 24a
14.3 Mb

654

References

Amemiya CT et al. 2013. The African coelacanth genome provides insights into tetrapod evolution. Nature. 496:311–316.

Amores A et al. 2014. A RAD-tag genetic map for the platyfish (*Xiphophorus maculatus*) reveals mechanisms of karyotype evolution among teleost fish. Genetics. 197:625–641.

Austin CM, Tan MH, Harrison KA, Lee YP, Croft LJ, Sunnucks P, Pavlova A, Gan HM. 2017. De novo genome assembly and annotation of Australia's largest freshwater fish, the Murray cod (*Maccullochella peelii*), from Illumina and Nanopore sequencing read. GigaScience. 6.

Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA. 6:11.

Barb JG, Bowers JE, Renaut S, Rey JI, Knapp SK, Rieseberh LH, Burke JM. 2014. Chromosomal evolution and patterns of introgression in helianthus. Genetics. 197:969–979.

Barrett RDH, Schluter D. 2008. Adaptation from standing genetic variation. Trends Ecol. Evol. 23:38–44.

Belaghzal H, Dekker J, Gibcus JH. 2017. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. Methods. 123:56–65.

Belton J-M, McCord RP, Gibcus, Naumova, Zhan Y, Dekker. 2012. Hi–C: A comprehensive technique to capture the conformation of genomes. Methods. 58:268–276.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.

Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods. 12:59–60.

Campagna L, Repenning M, Silveira LF, Fontana CS, Tubaro L. Pablo, and IJ Lovette. 2017. Repeated divergent selection on pigmentation genes in a rapid finch radiation. Science Advances. 3:e1602404.

Carneiro M, Rubin C-J, Di Palma F, Albert FW, Alföldi J, Martinez Barrio A, Pielberg G, Rafati N, Sayyab S, Turner-Maier J et al. 2014. Rabbit genome analysis reveals a polygenic basis for phenotypic change during domestication. Science. 345:1074–1079.

Catanach A, Crowhurst R, Deng C, David C, Bernatchez L, Wellenreuther M. 2019. The genomic pool of standing structural variation outnumbers single nucleotide polymorphism by threefold in the marine teleost *Chrysophrys auratus*. Mol. Ecol. 28:1210–1223.

Chaisson MJP, Sanders AS, Zhao X, Malhotra D, Porubsky D, Rausch T, Gardner EJ, Rodriguez OL, Guo L, Collins RL, et al. 2019. Multi-platform discovery of haplotype-resolved structural variation in human genomes. Nat. Commun. 10:1784.

Conover DO, Munch SB. 2002. Sustaining Fisheries Yields Over Evolutionary Time Scales.

Science. 297:94–96.

Conover DO, Arnott SA, Walsh MR, Munch SB. 2005. Darwinian fishery science: lessons from the Atlantic silverside (*Menidia menidia*). Canadian Journal of Fisheries and Aquatic Sciences. 62:730–737.

Conover DO, Kynard BE. 1981. Environmental sex determination: interaction of temperature and genotype in a fish. Science. 213:577–579.

Conover DO, Present TMC. 1990. Countergradient variation in growth rate: compensation for length of the growing season among Atlantic silversides from different latitudes. Oecologia. 83:316–324.

Corbett-Detig RB, Said, I, Calzetta M, Gdenetti M, McBroome J, Maurer MW, Petrarca V, della Torre A, Besansky. 2019. Fine-Mapping Complex Inversion Breakpoints and Investigating Somatic Pairing in the *Anopheles gambiae* Species Complex Using Proximity-Ligation Sequencing. Genetics. 213:1495-1511.

Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural selection constrains neutral diversity across a wide range of species. PLoS Biol. 13:e1002112.

Cunningham F, Achuthan P, Akanni W, Allen J, Amode MR, Armean IM, Bennett R, Bhai J, Billis K, Boddu S et al. 2019. Ensembl 2019. Nucleic Acids Res. 47:D745–D751.

Danecek P, Schiffels S, Durbin R. 2014. Multiallelic calling model in bcftools (-m).

De-Kayne R, Zoller S, Feulner PGD. 2020. A de novo chromosome-level genome assembly of *Coregonus* sp. 'Balchen': One representative of the Swiss Alpine whitefish radiation. Mol. Ecol. Resour. 20:1093-1109.

Dixon JR, Xu J, Dileep V, Zhan Y, Song F, Le VT, Yardimci GG, Chakraborty A, Bann DV, Wang Y, et al. 2018. Integrative detection and analysis of structural variation in cancer genomes. Nat. Genet. 50:1388–1398.

Dutoit L, Burri R, Nater A, Mugal CF, Ellegren H. 2017. Genomic distribution and estimation of nucleotide diversity in natural populations: perspectives from the collared flycatcher (Ficedula albicollis) genome. Mol. Ecol. Resour. 17:586–597.

Ellegren H. 2010. Evolutionary stasis: the stable chromosomes of birds. Trends Ecol. Evol. 25:283–291.

Faria R, Chaube P, Morales HE, Larsson T, Lemmon AR, Lemmon EM, Rafajlović M, Panova M, Ravinet M, Johannesson K et al. 2019. Multiple chromosomal rearrangements in a hybrid zone between Littorina saxatilis ecotypes. Mol. Ecol. 28:1375–1393.

Faust GG, Hall IM. 2014. SAMBLASTER: fast duplicate marking and structural variant read extraction. Bioinformatics. 30:2503–2505.

Fernandez-Silva I, Henderson JB, Rocha LA, Simison WB. 2018. Whole-genome assembly of the coral reef Pearlscale Pygmy Angelfish (*Centropyge vrolikii*). Sci. Rep. 8:1498.

Frith MC, Kawaguchi R. 2015. Split-alignment of genomes finds orthologies more accurately. Genome Biol. 16:106.

Götz S, García-Gómez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talón M, DopazoJ, Conesa A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res. 36:3420–3435.

Haenel Q, Laurentino TG, Roesti M, Berner D. 2018. Meta-analysis of chromosome-scale crossover rate variation in eukaryotes and its significance to evolutionary genomics. Mol. Ecol. 27:2477–2497.

Hao S, Han K, Meng L, Huang X, Shi C, Zhang M, Wang Y, Liu Q, Zhang Y, Seim I et al. 2020. Three genomes of Osteoglossidae shed light on ancient teleost evolution. bioRxiv. 2020.01.19.911958. doi: 10.1101/2020.01.19.911958.

Hice LA, Duffy TA, Munch SB, Conover DO. 2012. Spatial scale and divergent patterns of variation in adapted traits in the ocean. Ecol. Lett. 15:568–575.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-Genome Annotation with BRAKER. In: Gene Prediction: Methods and Protocols. Kollmar, M, editor. Springer New York: New York, NY pp. 65–95.

Holt C, Yandell M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 12:491.

Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay, Munson KM, Kronenberg ZN, Vives L, et al. 2017. Discovery and genotyping of structural variation from long-read haploid genome sequence data. Genome Res. 27:677–685.

Imakaev Mm Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. 2012. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat. Methods. 9:999–1003.

Jansen HJ, Liem M, Jong-Raadsen SA, Dufour S, Weltzien F-A, Swinkerls W, Koelewijn A, Palstra AP, Pelster B, Spaink HP et al. 2017. Rapid de novo assembly of the European eel genome from nanopore sequencing reads. Sci. Rep. 7:7213.

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S et al. 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 484:55–61.

Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, Yabana M, Harada M, Nagayasu E, Maruyama H et al. 2014. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 24:1384–1395.

Kess T, Bentzen P, Lehnert SJ, Sylvester EVA, Lien S, Kent MP, Sinclair-Waters M, Morris C, Wringe B et al. 2020. Modular chromosome rearrangements reveal parallel and nonparallel adaptation in a marine fish. Ecol. Evol. 10:638–653.

Kiełbasa SM, Wan R, Sato K, Horton P, Frith MC. 2011. Adaptive seeds tame genomic sequence comparison. Genome Res. 21:487–493.

Kim H-S, Lee B-Y, Han J, Jeong C-B, Hwang D-S, Lee M-C, Kang H-M, Kim D-H, Lee D, Kim J et al. 2018. The genome of the marine medaka *Oryzias melastigma*. Mol. Ecol. Resour. 18:656–665.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics. 5:59.

Lehmann R, Lightfoot DJ, Schunter C, Mitchell CT, Ohyanagi H, Mineta K, Foret S, Berumen ML, Miller DJ, Aranda M et al. 2019. Finding Nemo's Genes: A chromosome-scale reference assembly of the genome of the orange clownfish *Amphiprion percula*. Mol. Ecol. Resour. 19:570–585.

Lieberman-Aiden E, van Berkim Nl, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al. 2009. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 326:289–293.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecais G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 25:2078–2079.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760.

Lou RN, Fletcher NK, Wylder AP, Conover DO, Therkildsen NO, Searle JB. 2018. Full mitochondrial genome sequences reveal new insights about post-glacial expansion and regional phylogeographic structure in the Atlantic silverside (*Menidia menidia*). Mar. Biol. 165:124.

Lucek K, Gompert Z, Nosil P. 2019. The role of structural genomic variants in population differentiation and ecotype formation in *Timema cristinae* walking sticks. Molecular Ecology. 28:1224–1237.

Lu L, Zhao J, Li C. 2020. High-Quality Genome Assembly and Annotation of the Big-Eye Mandarin Fish (*Siniperca knerii*). G3. 10:877–880.

Machado AM, Tørresen OK, Kabeya N, Couto A, Petersen B, Felicio M, Campos PF, Fonseca E, Bandarra N, Lopes-Marques M et al. 2018. 'Out of the Can': A Draft Genome Assembly, Liver Transcriptome, and Nutrigenomics of the European Sardine, *Sardina pilchardus*. Genes. 9:485.

Marçais G, Kingsford C. 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 27:764–770.

Martinez Barrio A, Lamichhaney S, Fan G, Rafati N, Pettersson M, Zhang H, Dainat J, Ekman D, Höppner M, Jern P et al. 2016. The genetic basis for ecological adaptation of the Atlantic herring revealed by genome sequencing. eLife. 5:e12081

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 17:10–12.

Martin SH, Möst M, Palmer WJ, Salazar C, McMillan WO, Jiggins FM, Jiggins CD. 2016. Natural Selection and Genetic Diversity in the Butterfly *Heliconius melpomene*. Genetics. 203:525–541.

Mattingsdal M, Jentoft S, Tørresen OK, Knutsen H, Hansen MM, Robalo JI, Zagrodzka Z, André C, Gonzalez EB. 2018. A continuous genome assembly of the corkwing wrasse (*Symphodus melops*). Genomics. 110:399–403.

Mérot C, Oomen RA, Tigano A, Wellenreuther M. 2020. A Roadmap for Understanding the Evolutionary Significance of Structural Genomic Variation. Trends in Ecology & Evolution. 35:561-572

Miller JT, Reid NM, Nacci DE, Whitehead A. 2019. Developing a High-Quality Linkage Map for the Atlantic Killifish *Fundulus heteroclitus*. G3. 9:2851–2862.

Mollah MBR, Khan MGQ, Islam MS, Alam MS. 2019. First draft genome assembly and identification of SNPs from hilsa shad (*Tenualosa ilisha*) of the Bay of Bengal. F1000Res. 8:320.

Murray GGR, Soares AER, Novak BJ, Schaefer NK, Cahill JA, Baker AJ, Demboski JR, Doll A, Da Fonseca RR, Fulton TL et al. 2017. Natural selection shaped the rise and fall of passenger pigeon genomic diversity. Science. 358:951–954.

Mu Y, Huo J, Guan Y, Fan D, Xiao X, Wei J, Li Q, Mu P, Ao J, Chen X. 2018. An improved genome assembly for *Larimichthys crocea* reveals hepcidin gene expansion with diversified regulation and function. Commun Biol. 1:195.

Nguinkal JA, Brunner RM, Verleigh M, Rebi A, los Ríos-Pérez L, Schäfer N, Hadlich F, Stüeken M, Wittenburg D, Goldammer T. 2019. The First Highly Contiguous Genome Assembly of Pikeperch (*Sander lucioperca*), an Emerging Aquaculture Species in Europe. Genes. 10.

Ozerov MY, Ahmad F, Gross R, Pukk L, Kahar S, Kisand V, Vasemägi. 2018. Highly Continuous Genome Assembly of Eurasian Perch (*Perca fluviatilis*) Using Linked-Read Sequencing. G3. 8:3737–3743.

Pääbo S. 2003. The mosaic that is our genome. Nature. 421:409–412.

Pettersson ME, Rochus CM, Han F, Chen J. 2019. A chromosome-level assembly of the Atlantic herring genome—detection of a supergene and other signals of selection. Genome Res. 29:1919-1928

Policarpo M, Fumey J, Lafargeas P, Naquin D, Thermes C, Naville M, Dechaud C, Volff J-NCabau C, Klopp C et al. 2020. Contrasted gene decay in subterranean vertebrates: insights from cavefishes and fossorial mammals. bioRxiv. https://www.biorxiv.org/content/10.1101/2020.03.05.978213v1.abstract.

Putnam NH, O'Connell BL, Stites JC, Rice BJ, Blanchette M, Calef R, Troll CJ, Fields A, Hartley PD, Sugnet CW et al. 2016. Chromosome-scale shotgun assembly using an in vitro method for long-range linkage. Genome Res. 26:342–350.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26:841–842.

Rausch T, Zichener T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 28:i333–i339.

Reid NM, Proestou DA, Clark BW, Warren WC, Colbourne JK, Shaw JR, Karchner SI, Hanh ME, Nacci D, Oleksiak MF et al. 2016. The genomic landscape of rapid repeated evolutionary adaptation to toxic pollution in wild fish. Science. 354:1305–1308.

Reid NM, Jackson CE, Gilbert D, Minx P, Montague MJ, Hampton TH, Helfrich LW, King BL, Nacci DE, Aluru N et al. 2017. The landscape of extreme genomic variation in the highly adaptable Atlantic killifish. Genome Biol. Evol.

Robinson JA, Ortgea-Del Vecchyo D, Fan Z, Kim BY, vonHoldt BM, Marsden CD, Lohmueller KE, Wayne RK. 2016. Genomic Flatlining in the Endangered Island Fox. Curr. Biol. 26:1183–1189.

Rondeau EB, Minkley DR, Leong JS, Messmer AM, Jantzen JR, von Schalburg KR, Lemon C, Bird NH, Koop BF. 2014. The genome and linkage map of the northern pike (*Esox lucius*): conserved synteny revealed between the salmonid sister group and the Neoteleostei. PLoS One. 9:e102089.

Sardell JM, Cheng C, Dagilis AJ, Ishikawa A, Kitano J, Peichel CL, Kirkpatrick M. 2018. Sex Differences in Recombination in Sticklebacks. G3. 8:1971–1983.

Sardell JM, Kirkpatrick M. 2020. Sex Differences in the Recombination Landscape. Am. Nat. 195:361–379.

Sarropoulou E, Sundaram AYM, Kaitetzidou E, Kotoulas G, Gilfillan GD, Papandroulakis N, Mylonas CC, Magoulas A. 2017. Full genome survey and dynamics of gene expression in the greater amberjack *Seriola dumerili*. Gigascience. 6:1–13.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212.

Smit AFA, Hubley R. 2008. RepeatModeler Open-1.0. Available from http://www. repeatmasker. org.

Smit AFA, Hubley R, Green P. 2015. RepeatMasker Open-4.0. 2013--2015.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 24:637–644.

Stanke M, Schöffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 7:62.

Star B, Nederbragt AJ, Jentoft S, Grimholt U, Malmstrøm M, Gregers TF, Rounge TB, Paulsen J, Solbakken MH, Sharma A et al. 2011. The genome sequence of Atlantic cod reveals a unique immune system. Nature. 477:207–210.

Takehana Y, Zahm M, Cabau C, Klopp C, Roques C, Bouchez O, Donnadieu C, Brrachina C, Journot L, Kawaguchi M, et al. 2020. Genome Sequence of the Euryhaline Javafish Medaka, *Oryzias javanicus*: A Small Aquarium Fish Model for Studies on Adaptation to Salinity. G3. 10:907–915.

Tan MH, Austin CM, Hammer MP, Lee YP, Croft LJ, Gan HM. 2018. Finding Nemo: hybrid assembly with Oxford Nanopore and Illumina reads greatly improves the clownfish (Amphiprion ocellaris) genome assembly. GigaScience. 7: gix137.

Therkildsen NO, Wylder AP, Conover DO, Munch SB, Baumann H, Palumbi SR. 2019. Contrasting genomic shifts underlie parallel phenotypic evolution in response to fishing. Science. 365:487–490.

Therkildsen NO, Baumann H. 2020. A comprehensive non-redundant reference transcriptome for the Atlantic silverside Menidia menidia. Mar. Genomics. 100738.

Therkildsen NO, Palumbi SR. 2017. Practical low-coverage genomewide sequencing of hundreds of individually barcoded samples for population and evolutionary genomics in nonmodel species. Mol. Ecol. Resour. 17:194–208.

Tigano A, Colella JP, MacManes MD. 2020. Comparative and population genomics approaches reveal the basis of adaptation to deserts in a small rodent. Mol. Ecol.29:1300-1314.

Tigano A, Friesen VL. 2016. Genomics of local adaptation with gene flow. Mol. Ecol. 25:2144–2164.

Tigano A, Sackton TB, Friesen VL. 2018. Assembly and RNA-free annotation of highly heterozygous genomes: The case of the thick-billed murre (*Uria lomvia*). Mol. Ecol. Res. 18:79-90

Turner SD. 2014. qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots. bioRxiv. 005165. doi: 10.1101/005165.

Uwa H, Ojima Y. 1981. Detailed and Banding Karyotype Analyses of the Medaka, *Oryzias latipes* in Cultured Cells. Proc. Jpn. Acad. Ser. B Phys. Biol. Sci. 57:39–43.

Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 534:102–105.

Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowwski, Schatz MC.

2017. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 33:2202–2204.

Warkentine BE, Lavett Smith C, Rachlin JW. 1987. A Reevaluation of the Karyotype of the Atlantic Silverside, *Menidia menidia*. Copeia. 1987:222-224.

Weisenfeld NI, Kumar V, Shah P, Church DM, Jaffe DB. 2017. Direct determination of diploid genome sequences. Genome Res. 27:757–767.

Weissensteiner MH, Bunikis I, Catalán A, Francoijs K-J, Knief U, Heim W, Peona V, Pophaly S, Sedlazeck FJ, Suh A et al. 2020. Discovery and population genomics of structural variation in a songbird genus. Nat. Commun. 11:3403.

Wellenreuther M, Bernatchez L. 2018. Eco-Evolutionary Genomics of Chromosomal Inversions. Trends Ecol. Evol. 33:427–440.

Wilder AP, Palumbi SR, Conover DO, Therkildsen NO. 2020. Footprints of local adaptation span hundreds of linked genes in the Atlantic silverside genome. Evol Lett. 4:430–443.

Yuan Z, Liu S, Zhou T, Tian C, Bao L, Dunham R, Liu Z. 2018. Comparative genome analysis of 52 fish species suggests differential associations of repetitive elements with their living aquatic environments. BMC Genomics. 19:141.

Zhang D-C, Guo L, Guo H-Y, Zhu K-C, Li S-Q, Zhang Y, Zhang N, Liu B-S, Jiang S-G, Li J-T. 2019. Chromosome-level genome assembly of golden pompano (*Trachinotus ovatus*) in the family Carangidae. Sci Data. 6:216.